Formation
Logiciel R

Institut Pasteur
de Nouvelle-Calédonie

1

3° séance

DESCRIPTION DES DONNÉES AVEC LE LOGICIEL R

2

Types de variables

- **Quantitatives**
  - Continues (continuous data)
  - Discrètes (discrete data)
  - Temporelles
- **Qualitatives**
  - Ordinales (ordinal data)
  - Nominales (nominal data)
  - Binaires (binary or dichotomous data)

3

Variables quantitatives continues

Les valeurs de la variable varient de façon continue et sont exprimées avec une unité de mesure.

Exemples	
Poids	kg
Taille	m
Cholestérol	g/l
Pression artérielle	cm de hg

La précision de la mesure est le plus souvent limitée par l'instrument de mesure.

4

Variables quantitatives discrètes

La variable ne prend que quelques valeurs entières dénombrables, un nombre fini de valeurs.

Exemples

Congès annuels	Jours/an
Rappel de vaccins	Nbre d'injections
Parité	Nbre d'enfants
Activité de soin	Nbre de consultation

La transformation d'une variable quantitative continue en une variable quantitative discrète s'appelle la discrétisation ou groupement par classe (ex : classes d'âge).

5

Variables temporelles

La variable quantitative qui utilise les unités de mesure du temps

Exemples

Age de grossesse	semaine
Date de naissance	JJ/MM/AAAA
Durée d'hospitalisation	Jours
Délai d'intervention	Heures

Il existe des variables de durée et des variables servant à définir un instant donné (date) à partir duquel il est possible de calculer une durée.

6

Variables qualitatives ordinales

Une variable qui s'exprime en classes qui peuvent être ordonnées selon un échelle de valeurs

Exemples

Niveau d'étude	Prim., second., sup.
Complications d'une maladie	Modérée, moyenne, sévère
Score	De faible à élevé
Intensité d'une douleur	De faible à élevé

La codage numérique pour le traitement informatique ne doit pas les faire passer pour des variables qualitatives discrètes et n'autorise donc pas de le manipuler de façon arithmétique.

7

Variables qualitatives nominales

Les classes ne peuvent être hiérarchisées pour cette variable.

Exemples

Groupe sanguin	A,B,O, AB.
Etat civil	Célibataire, marié, divorcé…
Couleur des yeux	Bleu, vert, marron,…
Toiture	Tuile, tôle, paille…

L'ordre de présentation de ces variables est arbitraire.

8

Variables binaires

C'est un type particulier de variables nominales qui ne peuvent prendre que deux valeurs.

Exemples

Etat de santé	Malade, sain.
Sexe	Homme, femme.
Signe clinique	Présence, absence
Statut vaccinal	Vacciné, non vacciné

Autres appellations: variables dichotomiques, variables booléennes (vrai/faux), variables de Bernouilli codés 0 ou 1.

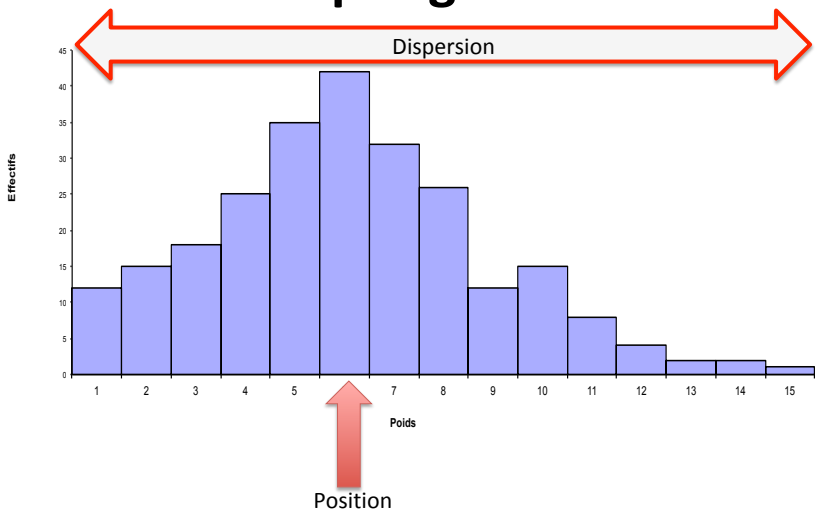
9

2° séance – 2° partie

DESCRIPTION DES DONNÉES AVEC LE LOGICIEL R

10

Principes généraux



11

STATISTIQUES DESCRIPTIVES

VARIABLES QUANTITATIVES

12

Paramètres de statistique descriptive

Paramètres de position	Fonctions R
Médiane	median(),summary()
Moyenne	mean(),summary()
Quartiles	quantile(),summary()

Paramètres de dispersion	Fonctions R
Minimum & étendue	min(),quantile(),summary()
Maximum & étendue	max(),quantile(),summary()
Variance	var()
Ecart-type	sd()
Intervalle de confiance	t.test()

13

Importer le fichier de travail

- Utiliser le fichier sauvegarder de la précédente journée
- Fonction read.table()

- Rappel
- Mydata<-read.table(file.choose(), header=T, sep= ";",dec=",", row.names="NUMIDENT")

14

Identifier les variables quantitatives

- On utilisera ici la fonction str()

'data.frame': 200 obs. of 17 variables:
$ NUMIDENT : Factor w/ 200 levels "AtR5678","AtR5679",..:
$ NAISSANCE : Factor w/ 119 levels "01/01/1928","01/01/1930",..: 9 2 ...
$ AGE : int 3919 915 1984 1515...
$ SEXE : int 2 1 2 2 2 2 2 1 1 2 ...
$ NIVSCOL : int 3 1 3 3 1 1 2 3 0 3 ...
$ NBHABFOYER: int 7 6 8 5 9 5 7 9 6 5 ...
$ DTEXAM : Factor w/ 53 levels "01/01/1 3 1 5 ...
$ VALEURF : num 37 38.1 38 38.9 38.4 37 40...
$ VOMISS : int 1 999 0 0 0 0 1 0 0 0 ...
$ CEPHAL : int 1 999 1 1 1 1 1 999 1 1 ...
$ ASTH : int 3 999 3 3 1 2 3 1 3 3 ...
$ DOULABD : int 0 999 0 0 0 0 0 0 0 0 ...
$ TOUX : int 1 999 1 0 0 0 0 0 1 0 ...
$ DIARR : int 1 1 999 0 0 0 0 1 0 0 ...
$ DTEDEBUT : Factor w/ 54 levels "01/01/2000",”3 ...
$ DTEFIN : Factor w/ 51 levels "","01/01/2000”, 8 ...
$ PALU : int 0 0 0 1 0 0 1 0 0 0 ...

15

Paramètres de position (3)

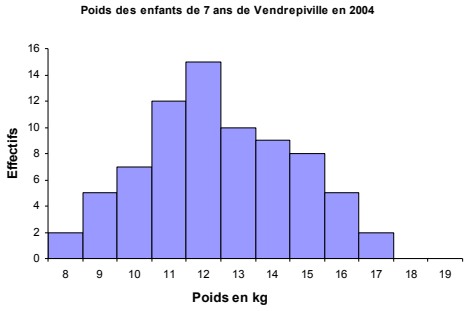
- **Moyenne** (mean)
  - Somme algébrique des valeurs observées divisée par le nombre de sujets

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

  - Paramètre de tendance centrale qui sert à résumer une série de donnée d'une variable quantitative

- **La moyenne est sensible au poids des valeurs extrêmes;**
  - Si dispersion homogène, la moyenne est un bon indicateur de distribution

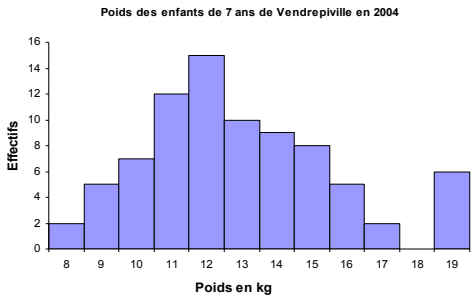
POIDS en Kg	NB enfants
8	2
9	5
10	7
11	12
12	15
13	10
14	9
15	8
16	5
17	2
18	0
19	0



Moyenne = $\sum n_i x_i / N$ = 933/75 =12,4

- **La moyenne est sensible au poids des valeurs extrêmes.**
  - S'il existe des valeurs extrêmes très élevées (basses ou hautes) la moyenne est un mauvais indicateur de tendance centrale

POIDS en Kg	NB enfants
8	2
9	5
10	7
11	12
12	15
13	10
14	9
15	8
16	5
17	2
18	0
19	6



Moyenne = $\sum n_i x_i / N$ = 1047/80 =13.08

Fonction mean()

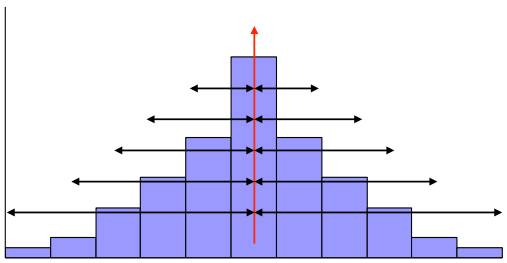
- Permet de calculer la moyenne d'une variable numérique

```
>mean(Mydata$VALEURF, na.rm=TRUE)

>mean(Mydata$AGEAN, na.rm=TRUE)

>mean(Mydata$NBHABFOYER, na.rm=TRUE)

#na.rm=TRUE signifie que les valeurs manquantes sont retirées avant
   le calcul
```

19

Paramètres de dispersion (2)

- **Variance**
  - Paramètre de dispersion le plus utilisé
  - Résume l'ensemble des écarts de chaque valeur d'une distribution par rapport à la moyenne



Paramètres de dispersion (3)

- **Variance : définition**
  - **C'est la moyenne des carrés des écarts à la moyenne**

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

  - **Variance : en pratique**

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

NB : Le dénominateur ici est N car on considère la série étudiée comme une population exhaustive

Paramètres de dispersion (4)

- **Variance : propriétés**
  - **Utilise toutes les valeurs de la distribution**
  - **Meilleur indicateur de dispersion d'une variable autour de la moyenne**
    - **Faible = dispersion resserrée**
    - **Elevé= dispersion élevée**

Inconvénient* : *s'exprime par une unité élevée au carré, ordre de grandeur différent de celui des valeurs de la distribution

Fonction var()

- Permet de calculer la variance d'une variable numérique

```
>var(Mydata$VALEURF, na.rm=TRUE)

>var(Mydata$AGEAN,na.rm=TRUE)

>var(Mydata$NBHABFOYER, na.rm=TRUE)
```

23

Paramètres de dispersion (5)

- **Ecart type** **(standard deviation)**
  - **C'est la racine carré de la variance**

$$\sigma = \sqrt{\sigma^2}$$

  - **Plus il est élevé, plus la dispersion est élevée**
  - **Plus il est faible, plus la dispersion est resserrée**

Avantage: ***il s'exprime dans la même unité que la moyenne***

Fonction sd()

- Permet de calculer l'écart type d'une variable numérique

```
>sd(Mydata$VALEURF, na.rm=TRUE)

>sd(Mydata$AGEAN,na.rm=TRUE)

>sd(Mydata$NBHABFOYER, na.rm=TRUE)
```

25

Intervalle de confiance de la moyenne

- Fonction t.test()

```
t.test(Mydata$VALEURF,conf.level=0.95) $conf.int

t.test(Mydata$AGEAN, conf.level=0.95) $conf.int

t.test(Mydata$NBHABFOYER, conf.level=0.95) $conf.int
```

26

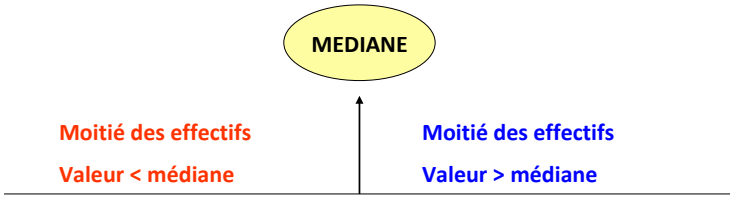
EXERCICE

```
> x <- rnorm(20,10)
# génère 20 nombres à partir d'une
#distribution normale de moyenne 10
> x
[1] 8.792221 9.585007 9.476082 9.766650 7.688314 7.452640 10.606369
[8] 10.221526 7.674439 10.097941 10.565285 7.831796 8.687138 10.393309
...
> t.test(x)
One Sample t-test
data : x
t = 35.5333, df = 19, p-value < 2.2e-16
alternative hypothesis : true mean is not equal to 0
95 percent confidence interval :
9.035537 10.166602
sample estimates :
mean of x
9.60107
```

27

Paramètres de position (1)

- **Médiane** (median)
  - **Valeur qui partage une série de données d'une variable quantitative en deux groupes d'effectifs égaux**



Fonction median()

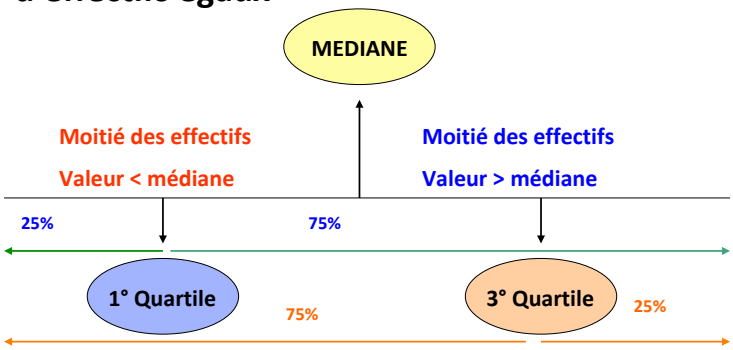
- Permet de calculer la médiane d'une variable numérique

```
>median(Mydata$VALEURF,na.rm=T)

>median(Mydata$AGEAN,na.rm=TRUE)

>median(Mydata$NBHABFOYER,na.rm=TRUE)
```

29

Paramètres de position (2)

- **Quartiles**
  - **3 valeurs qui partagent une série de données d'une variable quantitative en quatre groupes d'effectifs égaux**



Fonction quantile()

- Permet de calculer les valeurs de quartiles mais aussi la maximale et la minimale d'une variable numérique

```
>quantile(Mydata$VALEURF,na.rm=T)
>quantile(Mydata$AGEAN,na.rm=TRUE)
>quantile(Mydata$NBHABFOYER,
  na.rm=TRUE)
```

31

Paramètres de dispersion (1)

- **Extrêmes**
  - **Ce sont les deux valeurs extrêmes de la distribution.**
  - **Maximum et minimum**
- **Etendue**
  - **C'est la différence entre les deux valeurs extrêmes**
  - **En cas de valeurs aberrantes, l'étendue donne une fausse idée de la dispersion**
- **Intervalle interquartile**
  - **C'est la différence entre les valeurs du 1° et du 3° quartile**

NB : ces paramètres de dispersion sont associés à la médiane.

Fonction min()

- Permet de calculer la valeur minimale d'une variable numérique

```
>min(Mydata$VALEURF,na.rm=T)

>min(Mydata$AGEANna.rm=TRUE)

>min(Mydata$NBHABFOYER, na.rm=TRUE)
```

33

Fonction max()

- Permet de calculer la valeur maximale d'une variable numérique

```
>max(Mydata$VALEURF,na.rm=T)

>max(Mydata$AGEAN,na.rm=TRUE)

>max(Mydata$NBHABFOYER, na.rm=TRUE)
```

34

Fonction summary()

- Description d'une distribution
  (minimum, 1°quartile, médiane, moyenne,3° quartile, maximum)

  >summary(Mydata$AGEAN)

NB: cette fonction supprime directement les valeurs manquantes dans les calculs donc inutile de préciser na.rm=TRUE.

35

Décrire les variables quantitatives

- maliste<-c(2,5,7) # liste des données quantitatives

- summary(MyData[,maliste])

36

STATISTIQUES DESCRIPTIVES

VARIABLES QUALITATIVES

37

Paramètres de statistiques descriptives

- Fonction table()
- Fonction prop.table()
- Fonction margin.table()
- Fonction cumsum()

38

Fonction table()

- Permet de calculer les effectifs d'une variable qualitative ou quantitative discrétisé

```
>table(Mydata$SEXE)
```

39

Fonction prop.table()

- Permet de calculer les proportions d'une variable qualitative ou quantitative discrète
- **prop.table(x, margin=NULL) : donne les fréquences des cellules contenues dans le tableaux x.**
- L'argument **margin détermine par rapport à quelle marge du tableau les fréquences seront** calculées : **prop.table(x, margin=1) pour le total ligne et prop.table(x, margin=2) pour le total** colonne.

```
>prop.table(table(Mydata$SEXE))
```

40

Fonction cumsum()

- Permet de calculer les effectifs cumulés d'une variable qualitative ou quantitative discrètisée.

```
>cumsum(table(MyData$NIVSCOL))
```

41

Fonction margin.table()

```
>margin.table(x, margin=NULL) # donne la somme des entrées d'un tableau .
```

L'argument margin détermine pour quelle marge du tableau les sommes seront calculées margin.table(x, margin=1) pour le total ligne et margin.table(x, margin=2) pour le total colonne.

42

Intervalle de confiance d'une proportion

```
>prop.test(23,100)
1-sample proportions test with continuity correction
data : 23 out of 100, null probability 0.5
X-squared = 28.09, df = 1, p-value = 1.158e-07
alternative hypothesis : true p is not equal to 0.5
95 percent confidence interval :
0.1542150 0.3269410
sample estimates :
P = 0.23
```

43

Intervalle de confiance d'une proportion

```
>prop.test(table(Mydata$SEXE))
data:  table(Mydata$SEXE), null probability 0.5
X-squared = 1.125, df = 1, p-value = 0.2888
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4683438 0.6100945
sample estimates:
   p
0.54
```

44

Fonction summary()

- Permet de décrire les variables qualitative de la table de données, il faudra s'assurer qu'elles sont factorisées ou utiliser la fonction as.factor()

```
>summary(as.factor(Mydata$SEXE))
```

45

Construire une table de fréquence

```
Variable<-Mydata$NIVSCOL
Effectifs<-table(Variable)
Proportion<-round(prop.table(Effectifs)*100,digit=1)
E.cumul<-cumsum(Effectifs[c(4,3,2,1)])
P.cumul<-cumsum(Proportion[c(4,3,2,1)])

Nv1<-c(Effectifs[1],Proportion[1],E.cumul[4],P.cumul[4])
Nv2<-c(Effectifs[2],Proportion[2],E.cumul[3],P.cumul[3])
Nv3<-c(Effectifs[3],Proportion[3],E.cumul[2],P.cumul[2])
Nv4<-c(Effectifs[4],Proportion[4],E.cumul[1],P.cumul[1])

TablNv<-rbind(Nv1,Nv2,Nv3,Nv4)
colnames(TablNv)<- c("Eff","(%)","Eff Cum","(% cum)")
rownames(TablNv)<-c("NonScolar","Primaire","Secondaire","Universitaire")
TablNv
```

46

Résultats de la commande

	Eff	(%)	Eff Cum	(% cum)
NonScolar	43	21.5	200	100.0
Primaire	53	26.5	157	78.5
Secondaire	46	23.0	104	52.0
Universitaire	58	29.0	58	29.0

47

DESCRIPTION DES DONNEES EN VUE DE REALISER UNE ANALYSE UNIVARIÉE

VARIABLES QUANTITATIVES

48

STATISTIQUES DESCRIPTIVES

VARIABLES QUANTITATIVES

49

Fonction by()

- Permet l'étude d'une variable par strate

```
>by(Mydata$AGEAN, Mydata$SEXE,
mean,na.rm=T)
Mydata$SEXE: 1
[1] 3347.139
------------------------------------------------------------------------
Mydata$SEXE: 2
[1] 3191.804
```

50

Fonction by()

- Permet l'étude d'une variable par strate

>by(Mydata$AGEAN, Mydata$SEXE, summary)

```
Mydata$SEXE: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    109    1171    1803    3347    4618   14840
------------------------------------------------------------------------------
Mydata$SEXE: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    538    1022    2068    3192    3365   26340
```

51

Fonction tapply()

Applique une même fonction à une variable en fonction d'un ou plusieurs facteurs

>tapply(Mydata$VALEURF, list(Mydata$TOUX,Mydata$ASTH),mean,na.rm=T)

```
      0        1        2       3
0 38.05 38.57714 38.3913 39.01091
1 38.00 38.20000 38.3200 38.43077
```

52

Fonction aggregate()

Applique une même fonction à une variable en fonction d'un ou plusieurs facteurs

```
>aggregate(Mydata$VALEURF, list(Mydata$TOUX,Mydata$ASTH),mean,na.rm=T)
```

	Group.1	Group.2	x
1	0	0	38.05000
2	1	0	38.00000
3	0	1	38.57714
4	1	1	38.20000
5	0	2	38.39130
6	1	2	38.32000
7	0	3	39.01091
8	1	3	38.43077

53

DESCRIPTION DES DONNEES EN VUE DE REALISER UNE ANALYSE UNIVARIÉE

VARIABLES QUALITATIVES

54

Fonction table()

- Permet de calculer les effectifs d'une variable qualitative ou quantitative discrète

```
>table(Mydata$NIVSCOL,Mydata$SEXE)
```

55

Fonction prop.table()

- Permet de calculer les proportions d'une variable qualitative ou quantitative discrète
- **prop.table(x, margin=NULL) : donne les fréquences des cellules contenues dans le tableaux x.**
- L'argument **margin détermine par rapport à quelle marge du tableau les fréquences seront** calculées : **prop.table(x, margin=1) pour le total ligne et prop.table(x, margin=2) pour le total** colonne.

```
>prop.table(table(Mydata$NIVSCOL,Mydata$SEXE),margin=1)
```

➔Proportion en ligne

```
>prop.table(table(Mydata$NIVSCOL,Mydata$SEXE),2)
```

➔ Proportion en colonne

56

Fonction margin.table()

>margin.table(table(Mydata$NIVSCOL,Mydata$SEXE),1) # donne la somme des lignes entrées d'un tableau .

>margin.table(table(Mydata$NIVSCOL,Mydata$SEXE),2)

L'argument margin détermine pour quelle marge du tableau les sommes seront calculées margin.table(x, margin=1) pour le total ligne et margin.table(x, margin=2) pour le total colonne.

57

Construire une table de contingence

- Effectifs<-table(Mydata$NIVSCOL,Mydata$SEXE)
- Proportion<-round(prop.table(Effectifs,1)*100,1)
- Total<-margin.table(Effectifs,1)

- Table<-cbind(Effectifs[,1],Proportion[,1],Effectifs[,2],Proportion[,2],Total)

- colnames(Table)<- c("Male","(%)","Femelle","(%)","Total")
- rownames(Table)<-c("NonScolar","Primaire","Secondaire","Universitaire")
- Table

58

Résultats de la commande

	Male	(%)	Femelle	(%)	Total
NonScolar	27	62.8	16	37.2	43
Primaire	28	52.8	25	47.2	53
Secondaire	27	58.7	19	41.3	46
Universitaire	26	44.8	32	55.2	58

59